EDV-Konzept der Göttinger Arbeitsstelle des Deutschen Wörterbuchs

Stackmann, Karl; Schlaefer, Michael; Lenz, Anja; Miehe, Almut; Bader, Winfried

Historical Social Research, Vol. 19 — 1994 — No. 4, 87-100

HUMANITIES COMPUTING

EDV-Konzept der Göttinger Arbeitsstelle des Deutschen Wörterbuchs*

Karl Stackmann; Michael Schlaefer; Anja Lenz; Almut Miehe; Winfried Bader

Der Übergang des Deutschen Wörterbuchs ins Zeitalter der wissenschaftlichen Datenverarbeitung

Als ich mein Amt als Akademischer Leiter der Göttinger Arbeitsstelle des Deutschen Wörterbuchs vor gut 17 Jahren übernahm, waren Organisation und Arbeitstechnik noch ganz konventionell, sie unterschieden sich, zugespitzt gesagt, nicht wesentlich von dem, was in der Zeit der Brüder Grimm üblich war. Moderne Technik spielte keine Rolle - mit einer Ausnahme. Die Belegzettel des Wortarchivs für die Buchstaben D, E, F waren durch Rückkopieren aus einem verfilmten Quellencorpus von sehr großem Umfang hergestellt worden.

Beim Wechsel im Amt des Arbeits Stellenleiters machte Herr Bahr, an dessen Stelle im Jahre 1985 Herr Schlaefer trat, darauf aufmerksam, daß die Technik für die Herstellung von Belegzetteln aus den Filmen in naher Zukunft nicht mehr zur Verfügung stehen werde. Wenn die Filme ihren Wert für die weitere Arbeit an unserem nationalen Wörterbuch nicht verlieren sollten, müßte ungesäumt damit begonnen werden, die Kopien für die bisher nicht berücksichtigten Buchstaben des Alphabets herzustellen. Das nötigte Herrn Schlaefer, sich von Anfang an um die Einführung neuer Techniken zu bemühen. Die Versuche, ein thesaurusähnliches Zettelarchiv nach der Kopiermethode herzustellen, schlugen fehl. Es fand sich kein Geldgeber, der bereit gewesen wäre, die immensen Kosten zu übernehmen.

Im Zuge der verschiedenen Verhandlungen, die in diesem Zusammenhang zu führen waren, wurden wir immer wieder darauf hingewiesen, man müsse die Möglichkeiten der elektronischen Datenverarbeitung für die Zwecke des Wör-

---

*** Protokoll des 58. Kolloquiums über die Anwendung der EDV in den Geisteswissenschaften an der Universität Tübingen am 3. Juli 1993.**

terbuchs nützen. Dabei werde sich dann möglicherweise auch ein Weg für die Weiterbenutzung des Göttinger Corpus finden lassen. Es wurden daher erste Versuche mit dem Einbau von Personalcomputern in die Arbeitsabläufe angestellt. Sie führten zu Rationalisierungseffekten an einzelnen Stellen, folgten aber noch keinem Gesamtkonzept.

Eben zu der Zeit, in der diese Versuche liefen, wurde die Forderung der Geldgeber dringlicher, es müsse ein Konzept entwickelt werden, das eine Beendigung der Neubearbeitung D, E, F durch die Göttinger Arbeitsstelle bis zum Jahre 2005 gewährleiste. Dies war die Situation, als Herr Schlaefer und ich Ende Februar 1988 zum »4. Kongreß für Maschinelle Verarbeitung altdeutscher Texte« nach Trier fuhren. Die Vorstellung von TUSTEP, der wir dort beiwohnen konnten, überzeugte uns, daß mit diesem Programm die Probleme, vor denen wir standen, zu lösen sein würden. Wir fanden bei Herrn Ott und seinen Mitarbeitern freundliche Unterstützung. Mit ihrer Hilfe ist es in den letzten fünf Jahren gelungen, ein neues Gesamtkonzept zu entwickeln, das die Möglichkeiten der Datenverarbeitung voll für die Zwecke einer zeitgemäßen Lexikographie in Dienst nimmt. Alles was in den Arbeitsabläufen mit Hilfe der Datenverarbeitung rationalisiert werden kann, ist auf TUSTEP umgestellt. Wir können heute sogar die Druckvorlagen in der Arbeitsstelle anfertigen.

Karl Stackmann (Göttingen)

EDV-Konzept der Göttinger Arbeitsstelle des Deutschen Wörterbuchs

Das Deutsche Wörterbuch, das von den Grimms begonnen wurde, konnte nach 122 Jahren Bearbeitungsdauer im Jahr 1960 mit insgesamt 32 Bänden abgeschlossen werden. Das Werk repräsentiert das große sprachnationale historische Wörterbuch des Deutschen und damit ein Zentrum der deutschen Wörterbuchlandschaft.

Die Planung der Neubearbeitung wurde von Anfang an durch verschiedene Faktoren behindert. Zum einen wirkte sich die politische Teilung Deutschlands auf die Kooperationsmöglichkeiten der beiden Arbeitsstellen in Berlin (Ost) und Göttingen aus. Zum anderen entwickelten beide Arbeitsstellen bald eine beträchtliche Eigendynamik. Wohl mit Rücksicht darauf, nur mit ganz erheblichen Kompromissen die Neubearbeitung als ein deutsch-deutsches Unternehmen gestalten zu können, blieb es bei einer arbeitsorganisatorischen Lösung, nach der in Berlin die Buchstaben A-C und in Göttingen die Buchstaben D-F bearbeitet werden sollten. Konzeptionell beschränkte man sich auf allgemeine Zielvorgaben. Die Artikel sollten im Anschluß an die besten Traditionen des abgeschlossenen Werks erstellt werden. Beide Arbeitsstellen bauten unabhängig voneinander für ihre Buchstabenbereiche Zettelarchive von je etwa 2,5 Millionen Belegen auf. Die Drucklegung erfolgte bei VEB Maxim Gorki in Altenburg/Thüringen im Monotype-Bleisatz.

Historical Social Research, Vol. 19 — 1994 — No. 4, 87-100

Die Unternehmensentwicklung verlief bis zur Mitte der achtziger Jahre in beiden Arbeitsstellen zwar inhaltlich abweichend, jedoch strukturell vergleichbar. Die unscharfen konzeptionellen Vorgaben bewirkten vor allem in Verbindung mit dem Vollständigkeitspostulat einen sehr hohen Arbeitsaufwand für randständige Information und damit einen sehr langsamen Publikationsprozeß. In Berlin erschien bis 1985 etwa eine halbe Lieferung, in Göttingen etwa eine Lieferung pro Jahr. Hochrechnungen um 1985 zeigten, daß bei gleichbleibender Produktivität der Berliner Teil kaum vor 2070 und der Göttinger Teil etwa um 2020 abzuschließen sein würde. Benutzerreaktionen und kritische Bewertungen der Neubearbeitung aus wissenschaftlicher Sicht signalisierten Anzeichen eines möglichen Akzeptanzv**erlustes.**

In dieser Situation entstand ein Reorganisationsplan. Er koordinierte angestrebte Verbesserungen des Archivs, des Hilfsmittelbereichs und des Arbeitsprozesses. Darüber hinaus umschrieb er Umfangsbegrenzungen, thematische Schwerpunktbildungen und eine Zurücknahme des Vollständigkeitsanspruchs aus konzeptioneller Sicht. Zur Reorganisation gehörte auch die Einführung von Datenverarbeitungsansätzen vor allem im Hilfsmittelbereich.

Es wurde bald deutlich, daß punktuelle EDV-Lösungen eine arbeitsökonomisch und qualitätsbezogen wünschenswerte Ausschöpfung der insgesamt enthaltenen Möglichkeiten verstellten. Aus der Position »EDV« im Reorganisationskonzept mußte eine Position »EDV-Konzept« werden.

Gestützt auf die Erfahrungen der Orientierungsphase entstand ein EDV-Konzept, das zunächst Rahmenbedingungen für einen zusammenhängenden, unternehmensspezifischen Rechnereinsatz formulierte. Die Rahmenbedingungen des Konzepts sind seit ihrer ersten Umschreibung im Jahr 1988 im wesentlichen stabil geblieben. Es ist jedoch darauf zu verweisen, daß die Umsetzung des Rahmens in eine bestimmte EDV-Anwendungsstruktur erst in Auseinandersetzung mit TUSTEP als Philosophie und Software möglich war. Insofern stellt das Konzept heute eine Verbindung unserer Grundüberlegungen mit den durch TUSTEP gebotenen Datenverarbeitungsmöglichkeiten dar.

Ich komme zu den Rahmenbedingungen im einzelnen:

1. Datenverarbeitung muß projektspezifisch und problembezogen eingesetzt werden

Im Unterschied zu forschungsbezogenen Fragestellungen, wie Datenverarbeitung im lexikographischen Bereich eingesetzt werden kann, hat diese Technik in unserem Teil der Neubearbeitung nur die Funktion, bestimmte, fest umrissene Problemstellungen zu lösen. Der Datenverarbeitungseinsatz muß gegenüber alternativen Lösungen qualitativ bessere und/oder ökonomischere Ergebnisse ermöglichen. Dabei ist stets die Relation von Aufwand und Ergebnis zu beachten. Die Einführung der EDV darf den Ablauf der Neubearbeitung nicht stören.

2. Datenverarbeitung muß projektintern finanzierbar sein.

Unter den gegebenen Haushaltsbedingungen kommt längerfristig für die Göttinger Arbeitsstelle nur eine Ausstattung mit PCs in Frage. Diese Geräte können aus dem laufenden Sachhaushalt sukzessiv beschafft und betriebsbereit gehalten werden.

Aus dieser Rahmenbedingung ist ein dezentraler PC-Bestand in der Arbeitsstelle entstanden. Geringe Wartung, räumliche Flexibilität und wachsende Leistungsfähigkeit der Geräte haben sich bislang als Vorteile erwiesen. Da wir im Göttinger Neubearbeitungsteil erst zu einem sehr späten Zeitpunkt der Lieferungserstellung auf zentrale Daten zugreifen müssen, spielt das Vorhandensein einer Vernetzung nur eine sehr untergeordnete Rolle. Mit dem Übergang zur Arbeit an einem Historischen Wortschatzarchiv ändert sich diese Bedarfssituation. Dem wird durch die Installation eines kleinen PC-Netzverbandes Rechnung getragen.

3. Die Software muß philologischen und arbeitsökonomischen Bedingungen genügen.

Das für die Neubearbeitung benötigte komplizierte System von Sonderzeichen muß schon auf der Ebene der Typoskripterstellung verfügbar sein, damit keine Einschränkungen einer philologisch akzeptablen Textwiedergabe eintreten. Textverarbeitungsfunktionen müssen flexibel für verschiedene Zwecke einsetzbar sein. Lern- und Bedienungsaufwand sollen möglichst niedrig sein.

4. Verschiedene Softwarekomponenten sollen einfach kombinierbar sein.

Wenn Textverarbeitung und Datenbanken für die Neubearbeitung wirksam erschlossen werden sollen, müssen verschiedene Datenbanken und die Typoskripterstellung von den Programmen her kompatibel sein. Großer Konversionsaufwand ist aus Kapazitätsgründen zu vermeiden. Beim Übergang zum Computersatz ist ebenfalls auf die Kompatibilität der Satzprogramme mit Textverarbeitungs- und Datenbankprogrammen zu achten.

5. Datenverarbeitungsansätze müssen mit externer Beratung verbunden sein.

Da in der Arbeitsstelle keine Personalkapazität für eine hauptamtliche Beschäftigung mit EDV zur Verfügung steht, muß bei zusammenhängender Einführung von EDV eine angemessene externe Beratung gewährleistet sein.

Es bedarf keiner weiteren Hinweise, daß mit TUSTEP ein finanzierbares, philologisch einsetzbares, polyfunktionales Programm mit einem sehr wirkungsvollen Ausbildungs- und Beratungsumfeld zur Verfügung stand. Nach einer Demonstration in Trier 1988 wurde TUSTEP als Standardsoftware für die Ar-

beitsstelle etagefuhrt. Seine besonderen Vorteile lagen darin, daß es sich unseren Aufgabenstellungen anpassen ließ und nicht eine Anpassung des Unternehmens an bestimmte Programmdeterminanten verlangte. Konversions- und Kompatibilitätsprobleme entfielen, da Datenbanken, Textverarbeitung und Satz mit TUSTEP realisierbar waren.

Frau Lenz und später Frau Miehe absolvierten die TUSTEP-Ausbildung in Tübingen und wirkten innerhalb der Arbeitsstelle als Multiplikatoren. Als EDV-beauftragte wissenschaftliche Mitarbeiterinnen ist es heute vor allem ihre Aufgabe, die EDV-Anwendung auf verschiedenen Ebenen zu planen und zu realisieren. Daß dies bislang weitgehend reibungslos geklappt hat, verdanken wir neben dem Engagement der Damen vor allem der gleichbleibend freundlichen und kompetenten Beratung aus Tübingen. Wir haben uns bei sehr vielen Problemen unmittelbar auf die EDV-Kompetenz vor allem von Herrn und Frau Ott und Herrn Bader stützen können, so daß Lösungen, die aufgrund des Entwicklungsaufwandes für uns selbst kaum möglich gewesen wären, in kurzer Zeit zur Verfügung standen. Die Trennung in lexikographische und datentechnische Kompetenz hat eine sehr effektive Nutzung des Programms erschlossen, die dem Neubearbeitungsprojekt zugute kommt und vielleicht für andere Unternehmen Vorbildcharakter haben könnte. In besonderer Weise vorbildlich sind uns auch die EDV-gestützten lexikographischen Arbeiten von Herrn Sappler gewesen. Die Diskussionen mit ihm haben die Genese unseres Konzepts nachhaltig beeinflußt

Die Standardsoftware TUSTEP wird in der Göttinger Arbeitsstelle in den Bereichen Textverarbeitung, Datenbanken und Satzaufbereitung eingesetzt. Im weiteren sollen kurz die Grundstrukturen der drei Ebenen skizziert werden.

Der Datenbankbereich erweist sich als die umfangreichste Anwendungsebene. Sie ist organisatorisch differenziert in die Komplexe »Hilfsmittel« und »Historisches Wortschatzarchiv«. Als Hilfsmittel bezeichnen wir Datensammlungen, die arbeitsstellenintern zur Planung, Überprüfung oder Ergänzung von Artikeln und Wortstrecken dienen. Zu nennen sind hier etwa Stichwortlisten zum Wortbestand der abgeschlossenen Ausgabe des Grimmschen Wörterbuchs, zum Wortbestand des Göttinger Archivs und zu M. Heynes Deutschem Wörterbuch. Femer ein Wort- und Begriffsregister zum Buchstaben E, eine Bibliographie zur Wortforschung und eine Zusammenstellung aller Informationen über die 6000 in Göttingen für Belege ausgewerteten Quellen Besondere Hervorhebung verlangt eine Datenbank mit allen in Berlin und Göttingen verwendeten bibliographischen Standardnachweisen zu den Belegen Dieses Instrument setzt nicht nur eine vorhandene Liste auf Datenträger um, sondern faßt vier verschiedene Vorläuferkarteien zusammen.

Die Datenbank »Historisches Deutsches Wortschatzarchiv« zur sogenannten Sicherungsmaßnahme erhält den zeitweilig fraglich gewordenen Zugang zum Göttinger Quellenbestand für eine potentielle Weiterarbeit am Grimmschen Wörterbuch. Sie schafft u.a. die Möglichkeit, Thesauruskomponenten aus dem

Wörterbuch auszulagern und wird darüber hinaus mittelfristig auch Grundlagenmaterial für andere Forschungen zum Wortschatz und zur Einzelwortgeschichte bieten.

In einer sehr rudimentären Form existiert eine Lieferungsdatenbank. Sie umfaßt die maschinenlesbaren Teile der Neubearbeitung und bildet den Grundstock für metalexikographische Forschungsansätze.

Die Textverarbeitungsebene umfaßt vor allem die Erstellung der Lieferungen. Hier zeigt sich besonders die Wirksamkeit des Modells »eine Software für alle Anwendungen«. Mit Blick auf die Lieferungsdatenbank und die dort wünschenswerten Zugriffsmöglichkeiten und unter Berücksichtigung einer ökonomischen Satzaufbereitung wird der Artikeltext im Unterschied zur älteren Praxis konsequenter strukturiert Diese Veränderungen erlauben gleichzeitig eine organisatorische Flexibilisierung der Typoskripterstellung. An die Stelle der Volltyposkripterstellung durch eine Schreibkraft tritt ein arbeitsteiliges Baukastensystem, das u.a. eine Entzerrung von Kapazitätsengpässen möglich macht

Der Bereich des Computersatzes ist schon erwähnt worden. Die Erstellung der Lieferungsvorlagen im Baukastenprinzip wird durch die Strukturierung der Texte bereits an satztechnischen Anforderungen ausgerichtet. Im Unterschied zur Berliner Arbeitsstelle streben wir in Göttingen keine vollständige Satzaufbereitung in der Arbeitsstelle an.

Die Einführung der Datenverarbeitung in den Göttinger Neubearbeitungsteil des Deutschen Wörterbuchs kann mit den skizzierten Grundlinien im wesentlichen als abgeschlossen gelten. Die Entscheidung, TUSTEP als Standard-Programm für alle Anwendungen einzuführen, hat sich aus unserer Sicht bewährt und für das Projekt ausgezahlt. Ganz entscheidend dafür ist auch der Ausbildungs- und Beratungsrahmen für dieses Programm gewesen. Mit TUSTEP als Datenverarbeitungsgrundlage haben wir ein aus unserer Sicht einfaches, in sich geschlossenes EDV-System realisieren können. Es ist gelungen, die technische Rückständigkeit des Unternehmens zu durchbrechen, ohne daß philologische Belange durch äußere Zwänge unmittelbar berührt worden wären. Der Arbeitsprozeß ist mit Hilfe der EDV gründlicher durchstrukturiert und vor allem flexibilisiert worden. Die zur Arbeit mit TUSTEP erforderliche Kompetenz konnte gestuft entwickelt werden, so daß für die Mehrzahl der älteren Bearbeiter kein Technikschock eintrat. Aufgrund der beträchtlichen Investitions- und Entwicklungsaufwendungen ist der unmittelbare Beschleunigungseffekt der EDV-Einführung in das nach ganz anderen Vorgaben ausgerichtete Neubearbeitungsunternehmen zunächst noch relativ begrenzt geblieben.

In welchem Umfang sich die Investitionen arbeitsökonomisch rentieren, wird in den nächsten Jahren sichtbar werden. Die Rechtfertigung für diese Investitionen liegt nach unserer Auffassung schwerpunktmäßig ohnehin auf einer anderen Ebene.

Wörterbucharbeit, das zeigt die allgemeine Entwicklung der letzten Jahre ganz deutlich, ist mit den von der EDV erschlossenen, systematischen Arbeitsformen in der Lage, neue Qualitätsstandards zu erreichen und weiterführende Informationsangebote zu erschließen. Daß bei einem laufenden Unternehmen von der Struktur des Deutschen Wörterbuchs dabei Grenzen bestehen, ist nicht zu leugnen. Ein Verzicht auf Anpassungen des Deutschen Wörterbuchs an die neuen Standards würde die Position dieses Wörterbuchs in der deutschen Wörterbuchlandschaft und seine Rolle als wissenschaftlich-lexikographisches Leitprojekt nachdrücklich in Frage stellen. Wir glauben daher, mit der Einführung von TUSTEP und der darauf gestützten Erfassung und Strukturierung von unternehmensspezifischen Daten ein Stück Investition in die Zukunft des Deutschen Wörterbuchs erreicht zu haben.

Michael Schlaefer (Göttingen)

EDV-Einsatz in der Typoskripterstellung und in der Satzaufbereitung bei der Neubearbeitung des Deutschen Wörterbuchs

In dem folgenden Referat sind zentrale Elemente unseres EDV-Konzeptes vorzustellen, die die laufende Artikelarbeit und Lieferungserstellung betreffen. Es sind dies die arbeitsteilige Typoskripterstellung, die Auszeichnung durch Makros und die Benutzung der Datenbank »Zitiertitelbasiskartei«.

Ganz allgemein läßt sich für ein Wörterbuch sagen, daß ein Maximum an Informationen auf einem zulässigen Minimum an Raum darzustellen ist. Dabei sollten die Informationen gut und möglichst schnell zu verstehen sein. In der Regel gilt ein einheitliches Darstellungsverfahren für das gesamte Wörterbuch, so daß die Artikel in gewissem Grad vergleichbar sind und der Benutzer sich nicht von Artikel zu Artikel auf eine andere Form der Informationsdarbietung einstellen muß.

Diesen Forderungen nach Darstellungsökonomie, nach Verständlichkeit zugunsten einer raschen Zugriffsmöglichkeit und nach einer weitgehenden Homogenität versucht auch die Neubearbeitung des Deutschen Wörterbuchs (²DWB), u. a. durch die Strukturierung der Wörterbuchinformation, nachzukommen.

Wir unterscheiden in jedem Artikel drei, z. T. vier Artikelteile. Es handelt sich um Stichwortgruppe, Einleitungsteil, Bedeutungsteil und u. U. die Kompositionsgruppe.

Diese vier Artikelteile lassen sich wiederum in systematisch wiederkehrende Elemente zerlegen wie z. B. meta- und objektsprachliche Elemente, Gliederungsabschnitte und gegebenenfalls Gliederungsmarken, Beschreibungstext, Datierungen, Belegtexte in Belegreihen, bibliographische Nachweise für Belege (Zitiertitel), Kompositionsgruppenstichwort, Kompositionsgruppenwörter.

Die Tatsache, daß im DWB strukturierte Daten vorliegen, machen wir uns in der Göttinger Arbeitsstelle bereits bei der Eingabe der Artikel zunutze. Mit Hilfe von Makros, also festen Zeichenfolgen, kennzeichnen wir die wiederkehrenden Artikelelemente. Auf diese Weise bereiten wir eine erleichterte Lichtsatzaufbereitung vor und eröffnen Perspektiven für metalexikographische Auswertungen unseres Wörterbuchs.

Die Makros und ihre Aufgaben werden im Zusammenhang mit den weiteren Erläuterungen deutlich. Zunächst ist das Verfahren der arbeitsteiligen Typoskripterstellung kurz zu erklären.

Der Lexikograph legt als Ergebnis der Bearbeitung des ihm zugeteilten Belegmaterials das sogenannte Kastenmanuskript vor. Dieses enthält das der Artikelgliederung entsprechend sortierte Belegmaterial mit den jeweiligen lexikographischen Angaben.

Parallel dazu erstellt der Lexikograph am PC mit TUSTEP die Artikelstrukturübersicht. Er gibt jeweils Stichwortgruppe, Einleitungsteil, Gliederungsmarken, Beschreibungstexte, Datierungen und die entsprechenden Elemente der Kompositionsgruppen ein. Die Belege und die Zitiertitel werden bei der Eingabe zunächst ausgespart. Alle Artikelelemente der Strukturübersicht, die auch als Fassung A des Artikels bezeichnet wird, sind bereits mit den betreffenden Makros gekennzeichnet. Der Ausdruck dieser Datei liegt mit dem Kastenmanuskript der Arbeitsstellenredaktion zugrunde.

Erst wenn die lexikographische Bearbeitung und Redaktion abgeschlossen sind, wird die Datei vervollständigt Fassung B entsteht, indem die Belege zu den jeweiligen Datierungen geschrieben werden. Diese Datei wird um die Zitiertitel ergänzt und erst jetzt, mit der Fassung C, liegt ein vollständiger Artikel als TUSTEP-Datei vor.

Die Vorteile dieses arbeitsteiligen, schrittweisen Verfahrens liegen auf der Hand. Bei der Typoskripterstellung mit der Schreibmaschine mußten bereits für die Arbeitsstellenredaktion die Artikel vollständig geschrieben werden. Redaktionelle Änderungen zogen nicht selten die vollständig neue Abschrift von Artikeln nach sich. Diese zeitraubende Mehrfacharbeit wird nun gespart

Zu einem relativ frühen Zeitpunkt liegt der Entwurf eines Artikels vor, der im Unterschied zum vollständigen Kastenmanuskript schnell zu überblicken ist. Redaktionelle Absprachen, kollegiale Anregungen und Kritik sind flexibel zu handhaben und können leicht umgesetzt werden. Zum Teil können Kollegen oder der Arbeitsstellenleiter direkt in einer Kopie der Datei am PC Vorschläge erarbeiten. Durch den schrittweisen Aufbau der Typoskripte beziehen sich alle Veränderungen nur auf die Dateifassung A. Belege und Zitiertitel werden erst in die Datei eingefügt, wenn feststeht, daß sie tatsächlich gedruckt werden sollen.

Durch eine EDV-gestützte **Aufbereitung** der Standardquellennachweise für Belege sind weitere Arbeitserleichterungen möglich. Ursprünglich existierten vier verschiedene Zettelkarteien für Quellennachweise. Sie liegen nun zusammen in

Historical Social Research, Vol. 19 — 1994 — No. 4, 87-100
einer TUSTEP-Datei vor, der »Zitiertitelbasiskartei«. Rasche Zugriffe sind über Autor, Titel und ein internes Zählnummernsystem möglich, darüber hinaus aber auch über jede beliebige Suchzeichenfolge.

Ein identifizierter Quellennachweis kann aus der Datei nun direkt hinter den Beleg in der Artikelfassung C kopiert werden. Auf diese Weise sind bei der Typoskripterstellung in der Dateifassung C statt der vollständigen Eingabe nur Modifikationen an dem kopierten Titel erforderlich.

Trotz der deutlichen Vorteile der arbeitsteiligen Typoskripterstellung und der Kopiermöglichkeit der Zitiertitel fällt zusätzliche, vor allem organisatorische Arbeit an, etwa um den Ablauf der arbeitsteiligen Typoskripterstellung zu koordinieren.

Zusätzlicher Aufwand entsteht auch durch die bereits erwähnte Auszeichnung der Artikel durch Makros. Dieses ist jedoch vor dem insgesamt veränderten Produktionsprozeß der Lieferungen des DWB zu sehen

Bis 1989 wurde die Neubearbeitung im Bleisatz gedruckt. Dieses Herstellungsverfahren wäre mittlerweile finanziell völlig untragbar. Es galt also, rechtzeitig und möglichst vom Wörterbuch selbst ein realisierbares Drucklegungskonzept zu entwickeln. Dabei war zu berücksichtigen, daß wörterbuchexterne Personen, etwa aus einem Verlag oder einer Druckerei, so wenig wie möglich wörterbuchspezifische Kenntnisse benötigen sollten und daß ihnen überhaupt so weit wie möglich zugearbeitet werden müßte, um Aufwand und Kosten in einem vertretbaren Rahmen zu halten. Gleichzeitig war es ausgeschlossen, daß wir im Wörterbuch selbst die gesamte Entwicklung und weitere Ausführung der Drucklegung leisten könnten. Wir erzielten einen Kompromiß, der darauf basiert, daß die systematisch wiederkehrenden Artikelelemente mit Makros gekennzeichnet werden. Diese Makros sind sprechende Zeichenfolgen, die möglichst sparsam, aber als ausreichende Markierung für die Weiterverarbeitung gesetzt werden. Sie sind aufgrund ihrer direkten Entsprechung zu den üblichen Artikelelementen für Wörterbuchangehörige problemlos zu verwenden.

Anschließend können sie über das TUSTEP-Programm KOPIERE weiter umgesetzt und schließlich in Steuerzeichen für den hausinternen Ausdruck in Göttingen oder für die Lichtsatzaufbereitung umgewandelt werden.

Auf diese Weise ist es möglich, dem Verlag eine Diskette zu senden, die von der Wörterbucharbeitsstelle weitgehend, ohne spezielle Kenntnisse über Lichtsatz und die entsprechenden Programme zu benötigen, vorbereitet ist. Wir kooperieren über einen Kompromiß, der zur Zeit für den Verlag auch finanziell akzeptierbar ist und bei dem beide Seiten, das Wörterbuch und die lichtsatzvorbereitende Stelle, mit dem jeweils eigenen Spezialwissen arbeiten können.

Dadurch, daß die Daten strukturiert erfaßt und gekennzeichnet sind, eröffnen sich auch Möglichkeiten, gezielte Such- oder Kopiervorgänge einzuleiten. Dieses bietet die Chance, unsere eigene Arbeit effektiver analysieren zu können.

Anja Lenz (Göttingen)

Historical Social Research, Vol. 19 — 1994 — No. 4, 87-100

Dieser Beitrag wird sich mit dem Projekt »Sicherungsmaßnahme« als einem Beispiel für den Einsatz von EDV im Deutschen Wörterbuch beschäftigen. Die Sicherungsmaßnahme der Göttinger Arbeitsstelle gewährleistet den EDV-gestützten Zugang zum Quellenbestand der Göttinger Arbeitsstelle und dient mittelfristig dem Aufbau eines Wortschatzarchivs.

Das Göttinger Quellencorpus für den Neubearbeitungsteil D-F umfaßt 6.000 Texte des 8.-20. Jahrhunderts. Es bildet damit die derzeit größte Sammlung historischer deutscher Texte für ein lexikographisches Unternehmen.

Die Quellen sind zu Beginn der Neubearbeitung durch Mikroverfilmung erfaßt und durch fotomechanische Rückvergrößerung für die Erstellung eines Zettelarchivs verfügbar gemacht worden. Anfang der achtziger Jahre war es, durch technische Veränderungen bedingt, nicht mehr möglich, die Quellenfilme mit einem vertretbaren Kostenaufwand in Rückvergrößerungen umzusetzen. Das hätte zur Konsequenz gehabt, daß der Zugang zum Göttinger Quellencorpus verschlossen gewesen wäre. Dies hätte für lexikographische Unternehmen außerhalb der Neubearbeitung gegolten. Für die potentielle Weiterarbeit am Deutschen Wörterbuch wäre ein neues Quellencorpus nötig gewesen, und die historisch-philologischen Disziplinen hätten die einzigartig bestehende Grundlage für ein historisches Gesamtwortschatzarchiv des Deutschen verloren. Um den Zugang zum Quellencorpus offenzuhalten, wurde ursprünglich der Plan verfolgt, alle Filme vollständig rückzuvergrößern. Durch Kopie der Rückvergrößerung und durch Leseexzerption hätte Belegmaterial für die Fortführung über F hinaus gewonnen werden können.

Dieses Vorhaben ist aus zwei Gründen nicht umgesetzt worden. Zum einen konnte dieses Verfahren angesichts der EDV-Entwicklung als veraltet angesehen werden. Zum anderen zeigten Untersuchungen teilweise Mängel der Leistungsfähigkeit der Göttinger Materialgrundlage auf. So konnten einerseits Beleglücken, auf der anderen Seite aber auch Belegredundanzen bei der Bezeugung von Stichwörtern nachgewiesen werden. Der zusätzliche Aufwand, der bei der Arbeit mit einem solchen Belegarchiv notwendig entsteht, wäre bei einer Reproduktion des Filmarchivs weiter tradiert worden. Außerdem hätte man sich bei der Gewinnung von Belegmaterial wieder auf die unsicheren Ergebnisse der Leseexzerption verlassen müssen, was letztlich die Erweiterung des alten Zettelarchivs bedeutet hätte.

Aus diesen Gründen kam es zur Neukonzeption des Projektes auf EDV-Grundlage. Trotz einiger Nachteile, die im Zeit- und Personalaufwand sowie im Speicherbedarf für die Daten liegen, hat dieses Verfahren große Vorteile. So eröffnen sich durch eine Digitalisierung größere Möglichkeiten der Weiterverarbeitung der Texte und, vor allem, durch eine EDV-gestützte Exzerption die Möglichkeit des planvollen und flexiblen Archivaufbaus.

Ich möchte kurz einige grundlegende konzeptionelle Festlegungen der Sicherungsmaßnahme umreißen.

Elektronisch gespeicherte Texte erschließen ein erheblich dichteres Stichwortnetz auf der Ebene des konventionellen Sprachgebrauchs, so daß die Gefahr der Anhäufung von Belegmassen besteht Abgesehen von technischen Problemen wäre der Großteil dieses Materials auch aus lexikographischer Sicht wertlos, da nur Massenbezeugungen typgleicher Vorkommen vorlägen. Um dem vorzubeugen, sind verschiedene Festlegungen getroffen worden:

Die erste Festlegung ist die der Zweckgebundenheit des Projektes. So dient die Sicherungsmaßnahme vorrangig dem Aufbau eines elektronisch gespeicherten Wortschatzarchivs A-Z, mit dem Material für die laufende Neubearbeitung bereitgestellt werden soll und das als Archiv für die potentielle Weiterarbeit am Deutschen Wörterbuch dienen kann.

Die zweite Festlegung besagt, daß zwischen der Texterfassung und der Textauswertung eine Interdependenz bestehen soll. Abgesehen davon, daß nur ein Teil der 6.000 Quellen digitalisiert werden soll, regelt diese Festlegung, daß das Textvolumen an das Exzerptionsergebnis gekoppelt werden soll. Und zwar wird es von der Qualität der Abbildfunktion der Texte für wort- und wortschatzgeschichtliche Sachverhalte abhängig gemacht.

Das genannte Ziel, das Belegarchiv planvoll und flexibel aufzubauen, erfordert eine Reihe von aufeinander abgestimmten Modellen, so etwa stark modellierte Vorstellungen vom Objektbereich des Deutschen Wörterbuchs, ein Texterfassungsmodell, mit dem man organisiert auf die Texte zugreift, sowie ein Exzerptionsmodell. Auf diese Modelle kann in diesem Zusammmenhang nicht näher eingegangen werden. Vielmehr möchte ich zu Beispielen für die konkrete Anwendung von TUSTEP in der Sicherungsmaßnahme kommen und dabei zunächst den Schritt der Digitalisierung der ausgewählten Quellen erläutern.

Die Art der Digitalisierung ist von der drucktechnischen Qualität der Texte abhängig. Wenn die Qualität eines Textes es gestattet, wird er mit dem Scanner (KDEM, OPTOPUS) automatisch digitalisiert und mit TUSTEP weiterverarbeitet Drucktechnisch schwierige Texte werden abgetippt.

Die derzeitige Planung für die Texterfassung umfaßt je 25 Texte der Textschienen literarische Texte, Rechtstexte und Periodika des Zeitraumes von ca. 1400 - ca. 1970. Die erfaßte Gesamtseitenzahl pro Textschiene beträgt ca. 5000 Buchseiten, d. h. ca. 200 pro Text. Bislang sind insgesamt ca. 30 Texte erfaßt worden. Damit ist die Voraussetzung gegeben, erste Messungen der lexikalisch-wortgeschichtlichen Leistungsfähigkeit des Corpus vorzunehmen. Dazu wurde ein Minitextcorpus gebildet. Aus den drei genannten Textschienen wurden jeweils 6 Texte ausgewählt, die für die Exzerption und anschließende Experimente zur Artikelerstellung für Einzelwörter aufbereitet werden sollen.

Die Exzerption wird gegliedert in eine makrostrukturelle Exzerption, in der es um die Ermittlung des Stichwortbestandes und der Stichwortstruktur der erfaß-

Historical Social Research, Vol. 19 — 1994 — No. 4, 87-100

ten Texte geht, und eine mikrostrukturelle Exzerption, die die Auswahl von Einzelbelegen zu einem Stichwort betrifft. Das angestrebte Ergebnis der makrostrukturellen Exzerption ist einmal eine Stichwortdatenbank als Verwaltungsinstrument zur Erschließung der Stichwortstruktur der erfaßten Texte mit den Informationen: Lemma, orthographische und flexivische Varianten, Textnachweis, Häufigkeit des Vorkommens. Außerdem entsteht ein Hintergrundsarchiv, das aus den lemmatisierten Konkordanzen besteht. Diese bieten ein erstes, wenn auch noch nicht weiter selektiertes, Belegmaterial.

Im Moment sind wir dabei, aus den erfaßten Texten Wortformenkonkordanzen (KWIC-Indizes) zu erstellen und diese zu lemmatisieren. Die ersten vier Konkordanzen wurden ohne Computerunterstützung lemmatisiert Sie bilden die Basis für die Stichwortdatenbank. Gegenwärtig wird jedoch auch mit EDV-unterstützter Lemmatisierung experimentiert. Dabei werden die nicht lemmatisierten Konkordanzen mit der genannten Stichwortdatenbank verglichen. Maschinell identifizierte Formen erhalten einen Lemmavorschlag, über den dann der Lemmatisierer entscheidet. Die bislang erreichten »automatischen« Erkennungen liegen bei etwa 30% der Textwörter. Eine Verbesserung der Erkennungsquote wird sich mit wachsendem Lemma- und Wortformenbestand ergeben. Dazu wird auch die Integration der ca. 350.000 Stichwörter der abgeschlossenen Ausgabe des Deutschen Wörterbuchs in die Stichwortdatenbank beitragen.

Man kann sich die Frage stellen, wieso nicht versucht wurde, allgemeine morphologische Regeln (etwa bei der Sortierung) anzugeben, also etwa nach Wortstämmen zuzuordnen.

Einmal sind die Wortstämme nicht immer klar zu isolieren, weil die Affixe selbst oft mehrere Bedeutungen haben. Zum anderen ergibt sich das Problem, daß wir es mit einem Corpus historischer Texte zu tun haben, das noch nicht einmal »historisch« homogen ist, sondern in dem die Texte sich über mehrere Epochen erstrecken. Die erforderliche Regelhaftigkeit im grammatischen, besonders aber im orthographischen Bereich, fehlt also.

Ein anderes, sprachimmanentes Problem sind z. B. die unfesten Verbzusammensetzungen. Sie lassen sich nicht computer-gestützt lemmatisieren, ohne daß der Text vorher mit Kodierungen aufbereitet wird. Da unser Ziel aber nicht vorrangig eine strukturierte Textdatenbank ist, sondern ein Belegarchiv, erschien es uns sinnvoller, die Lemmatisierung gleich an Konkordanzen vorzunehmen.

Nach den bisherigen Planungen soll das Wortschatzarchiv ab 1995 benutzbar sein.

Almut Miehe (Göttingen)

1. Typoskripterstellung und Satzaufbereitung

Das Konzept, das der Arbeit mit Makros beim Arbeitsprozeß der Typoskripterstellung zugrundeliegt, ist das der sachlichen Auszeichnung - im Gegensatz zu einer Beschreibung der Typographie und des Layouts - unter Verwendung von eindeutigen, mnemonisch gewählten Zeichenfolgen. Ein solches Konzept ist unter dem Gesichtspunkt des ungehinderten Datenaustausches unter dem Namen SGML (Standard Generalized Markup Language) als ISO-Norm festgeschrieben. In der Göttinger Arbeitsstelle wird das Verfahren aber unter dem Gesichtspunkt der Arbeitsteilung eingesetzt: der/die Bearbeiter/in soll den Wörterbuchartikel im Computer so vor sich haben, wie er/sie ihn sich denkt. Diese - durch die Redaktion vorgegebene - inhaltliche Struktur wird beim Erfassen des Artikels im Computer mit eingegeben. Dies verhindert zudem, daß notwendige Strukturteile vergessen werden.

Die Auszeichnungen, die in den Text eingegeben werden müssen, wurden auf ein Minimum reduziert: ein kleines Inventar von Auszeichnungen, keine Endemarken, Zeilenwechsel (der bei der Eingabe intuitiv richtig erfolgt) in der Datei ersetzen teilweise explizite Auszeichnungen.

Die Verbindung zur anderen Seite, der an vollständiger typographischer Information interessierten Weiterverarbeitung zum Lichtsatz und - ein Vorstadium dazu - dem Ausgeben von Kontrollausdrucken in der Arbeitsstelle für Korrekturzwecke, leistet ein Programm. Dieses Programm wandelt die minimale Auszeichnung der Eingabe *(minimized tagging*) um in eine vollständige und differenzierte Form, bei der alle Auszeichnungen auch mit Endemarken versehen sind, alle Auszeichnungen eindeutig gemacht werden und keine Information mehr nur implizit enthalten ist Dies geschieht durch algorithmische Verarbeitung, bei der Hierarchien und damit verschiedene Bedeutungen von gleichen Auszeichnungen erkannt, die Auszeichnungen durch ihre Umgebung unterschieden und die Zeilenwechsel berücksichtigt werden. Dieses Zwischenformat kann durch ein einfaches Ersetzen der Auszeichnungen in typographische Steueranweisungen für den Lichtsatz oder für die Kontrollausdrucke überführt werden Gleichzeitig dient das Programm dem Plausibilitätstest auf Vollständigkeit und Richtigkeit der Eingabecodierungen.

Das besondere an diesem Verfahren gegenüber der gängigen Verwendung von Makros für die Lichtsatzaufbereitung ist, daß die Auszeichnungsarbeit sich auf ein absolutes Minimum reduziert und die Arbeit des vollen Auszeichnens einem Programm überlassen wird. Das erzeugte Zwischenformat fügt sich in das modulare Konzept. Kann und muß der/die Bearbeiter/in nur das Eingabeformat kennen, so erfordert andererseits der Lichtsatz nur noch die Kenntnis der bereits erzeugten typographischen Steueranweisungen. Das Zwischenformat dagegen - zwar durch ein extern geschriebenes Programm erzeugt - ist auch für die EDV-Spezialistinnen in der Arbeitsstelle lesbar und kann von

ihnen mit einfachen Programmen zu verschiedenen Ausdruckformaten verarbeitet werden. Die vorgestellte Lösung ist auch eine Antwort auf das Problem des Einsatzes von externen Programmen, ohne daß diese wie eine Blackbox unzugänglich bleiben.

2. Erstellung des historischen Wortschatzarchivs

In diesem Bereich liegt die Aufgabe der EDV darin, das Erstellen des Belegarchivs mit Hilfe von KWIC-Indizes, die zu den einzelnen Texten des Textcorpus erstellt werden, zu erleichtern. Für die notwendige Lemmatisierung dient die EDV als Hilfsmittel mit der Aufgabe, notwendige Handarbeit - diese wird nicht durch Lemmatisierungsalgorithmen wegprogrammiert - zu minimalisieren, insbesondere jeden Arbeitsschritt nur genau einmal ausführen zu lassen, und die Verwaltung der von Hand lemmatisierten Belege zu übernehmen. Ist in einem KWIC-Index einmal von Hand die Lemmatisierung eingetragen, so geschieht das Zusammensortieren der Belege zu einem Lemma, das Eingliedern des Lemmas mit seinen Wortformen und der jeweiligen Herkunft in die Stichwortdatenbank, das Abspeichern aller Belege eines Textes in einem gesonderten Archiv sowie die Vorbereitung der Auswahl von Belegen für das Belegarchiv durch quantitative Auswertung der gesamten Belege automatisch.

Die Stichwortdatenbank (zu jedem Lemma sind hier sämtliche belegte Wortformen aufgeführt) dient auch als Grundlage für ein Programm, welches bei der Bearbeitung neuer Text automatisch Vorschläge für eine Lemmatisierung macht, so daß sich die Handarbeit ständig reduziert und sich mehr und mehr auf die Problemfälle und Homographen beschränken kann.

Winfried Bader (Tübingen)